Entwicklung eines dynamischen Entry Vocabulary Moduls für die Stiftung Wissenschaft und Politik

Benjamin Berghaus, Michael Kluck und Thomas Mandl

Universität Hildesheim
Informationswissenschaft
Marienburger Straße 22
31134 Hildesheim
benjamin.berghaus, mandl@uni-hildesheim.de

Stiftung Wissenschaft und Politik
Fachinformationsbereich
Ludwigkirchplatz 3-4
10719 Berlin
michael.kluck@swp-berlin.org

Abstract

Nicht übereinstimmendes Vokabular zwischen Anfrage und Dokumenten stellt ein Hauptproblem im Information Retrieval dar. Das Entry Vocabulary Modul hat sich in den letzten Jahren als Lösung hierfür etabliert. In diesem Beitrag wird ein dynamisches Entry Vocabulary Modul vorgestellt, das für einen Datenbestand mit mehreren inhaltsbezogenen Feldern in einem mehrstufigen Verfahren abhängig von Zwischenergebnissen die Anfrage erweitert. Das entwickelte System wurde anhand eines mehrsprachigen Datenbestands von rund 600.000 Fachtexten evaluiert und führte zu positiven Ergebnissen.

1 Einleitung

1.1 Die zentrale Frage des Vokabulars

Bei der Verbalisierung von Informationen wird die Nachricht mit Hilfe eines Vokabulars kodiert. Da es aufgrund verschiedener Sprachen und spezialisierter Fachsprachen viele verschiedene Vokabulare gibt, ist es essentiell, dass, sofern die Information ausgetauscht werden soll, sowohl der Sender als auch der Empfänger der Information das selbe Vokabular beherrschen und den Sinn der verbalen Abbildung der Information verstehen können. Ist das in der Kommunikation verwendete Vokabular einem der Kommunikationspartner unbekannt, wird der Austausch von Informationen nahezu unmöglich.

Bezogen auf die Welt des Information Retrieval ergibt sich in diesem Kontext ein ähnliches Problem. Je nach Aufgabe und Einsatzgebiet des IR-Systems variiert die Art der Datengrundlage und der in der Datengrundlage verzeichneten Informationen drastisch. Handelt es sich um eine hochspezialisierte Datenbank, beispielsweise die in [Gey *et al.*, 2001] herangezogene Datenbank von amerikanischen Im- und Exportstatistiken, so wird auch die Information in der Datengrundlage entsprechend in einem spezialisierten Vokabular kodiert sein. Im Falle der Außenhandelsstatistiken lässt sich beispielsweise nicht erfolgreich mit dem Begriff „automobile“ suchen - der entsprechende Begriff lautet in dem Zielvokabular des IR-Systems „Pass Mtr Veh“, was einen Abkürzung für „Passenger Motor Vehicle“ darstellt.

Üblicherweise würde eine solche Datengrundlage nur für die entsprechenden Spezialisten interessant sein, die mit dem entsprechenden Vokabular der Datengrundlage vertraut sein müssen. Allerdings kann es sein, dass auch eine solche Datenbank öffentlich zugänglich gemacht wird und somit auch Nutzern durchsucht wird, die sich des speziellen Vokabulars nicht bewusst sind. Hierbei entsteht, wie im oben beschriebenen Beispiel, die Situation, dass Nutzer, die des Vokabulars des Systems nicht mächtig sind, das System nicht auf eine zielführende Art und Weise bedienen können - nicht nur, weil sie das Ergebnis des Retrievalprozess eventuell nicht interpretieren, sondern weil auf das System unvorbereitete Nutzer ohnehin kaum eine sinnvolle Anfrage formulieren können.

Für die Lösung dieser Probleme der semantischen Heterogenität in Metadatensystemen existieren mehrere Ansätze, vgl. [Hellweg et al., 2001]. Um eine Brücke zwischen dem spezialisierten, kontrollierten Vokabular einer spezialisierten Datengrundlage und dem mehr oder weniger freien Vokabular eines untrainierten Nutzers zu bauen, wurden in den letzten Jahren zunehmend sogenannte Entry Vocabulary Module eingesetzt, vgl. [Buckland *et al.*, 1999]. Diese Module bestehen üblicherweise aus einem Entry Vocabulary Index, der die Beziehungen zwischen Termen des Freitexts und Deskriptoren oder Klassifikationsangaben auf Basis von Wahrscheinlichkeiten abbildet und einer Schnittstelle, die geeignete kontrollierte Vokabeln vorschlagen kann, vgl. [Norgard, 1998]. Auf diese Art und Weise kann eine Anfrage, die frei formuliert wurde, auf das eventuell kontrollierte Vokabular der Datengrundlage übersetzt oder um verwandte Terme oder Phrasen ergänzt werden.

Eine weitere, interessante Anwendungsmöglichkeit besteht außerdem darin, nicht nur einen „vertikalen“ Vokabularunterschied zu nivellieren, sondern auch einen „horizontalen“: Während der Unterschied zwischen spezialisiertem und freien Vokabular eindeutig ist, ist auch der Unterschied zwischen dem Vokabular verschiedener Sprachen - also der mehrsprachige Aspekt - durch den Einsatz von EVMs gegebenenfalls zu überbrücken. In [Petras, 2005] wurde bereits belegt, dass mehrsprachiges Information Retrieval durch den Einsatz von Metadaten verbessert werden kann: Petras wendete das EVM für die mit Thesaurustermen indexierte Fachdatenbank GIRT (German Indexing und Retrieval Testdatabase) an. GIRT wird zur Evaluie-

rung von mehrsprachigen Information Retrieval Verfahren im Rahmen des Cross Language Evaluation Forum[1] eingesetzt, vgl. [Mandl, 2006].

Neben der Übersetzung der eingegebenen Suchanfrage ist darüber hinaus deren Ergänzung um verwandte Terme und Phrasen möglich - hierbei steht nicht unmittelbar im Vordergrund, vollkommen verschiedene Vokabulare zu verknüpfen und somit überhaupt eine Suche möglich zu machen, sondern vielmehr die Retrievalleistung einer Anfrage zu verstärken.

1.2 Stiftung Wissenschaft und Politik

Die zugrundeliegenden Daten, für die das Information Retrieval System entwickelt und auf denen es evaluiert wird, werden von der Stiftung Wissenschaft und Politik, Berlin, (SWP) zur Verfügung gestellt. Es handelt sich um einen umfassenden Auszug aus der Literaturdatenbank des Fachinformationsverbunds für Internationale Beziehungen und Länderkunde (FIV). Die Literaturdatenbank ist beispielsweise über [Virtuelle Fachbibliothek Politikwissenschaften, 2006] zu nutzen, umfassende Informationen finden sich unter [Fachinformationsverbund IBLK, 2006].

Die SWP wurde 1962 bei München gegründet, hat ihren Hauptsitz seit 2001 in Berlin und ist ein deutsches Institut im Forschungsfeld der Außen- und Sicherheitspolitischen Fragen. Wichtigste Auftraggeber der SWP sind der Deutsche Bundestag, die Bundesregierung und die Ministerien, vorrangig hierbei das Auswärtigen Amt und das Verteidigungsministerium.

Die Datenbank des FIV umfasst rund 600.000 Einträge, die sich zum Großteil mit den Themengebieten Staat- und Gesellschaft, nationale und internationale Wirtschaft, Internationale Politik und Sicherheit befassen. Geographisch beziehen sich mehr als die Hälfte der verzeichneten Dokumente auf Europa, europäische Organisationen und die NATO. Weitere wichtige und berücksichtigte geographische Regionen schließen Afrika und den Nahen Osten, Nord- und Südamerika und Asien und Ozeanien neben anderen mit ein.

Die Datengrundlage verzeichnet zu 65% Bücher und Paper, zu 25% monographische Veröffentlichungen und zu jeweils 5% Periodika und Jahrbücher sowie Amtliche Veröffentlichungen. 24% der zu Verfügung gestellten Dokumente beinhalten ein Abstract, die restlichen Dokumente verfügen ausschließlich über einen Titel und ggf. diverse Deskriptoren als Metainformationen. Sprachlich dominieren die englischen Dokumente mit 51% die Datengrundlage. Deutsche Dokumente machen 28% des Umfangs aus, französische rund 11% und spanische 5%, während der Rest der Dokumente in sonstigen Sprachen verfasst ist. Die Deskriptoren sind insgesamt auf Deutsch verfasst. [Stiftung Wissenschaft und Politik, 2006]

2 Konzeption eines Entry Vocabulary Moduls

Das Konzept der Entry Vocabulary Modul wurde unter anderem in [Gey *et al.*, 2001] vorgestellt. In dieser Arbeit wurde die vier zentralen Komponenten wie folgt bezeichnet:

- eine ausreichend große Datengrundlage zum Trainieren des Entry Vocabulary Index
- ein Part-of-Speech-Tagger, der Substantive aus Dokumenttexten extrahiert
- ein Algorithmus, der die Beziehung zweier Begriffe anhand der Wahrscheinlichkeit ihrer Kooexistenz in einem Dokument errechnet
- das grundlegende Retrievalsystem, das die Suchanfrage entgegen nimmt und die Ergebnisse auflistet

Das von Gey vorgestellte System ist ein globaler (d.h. auf den gesamten Datenbestand bezogener) Ansatz zur Konstruktion eines Entry Vocabulary Index, in dem Terme des freien Vokabulars mit den kontrollierten Vokabeln der Metadaten auf Basis von probablistischen Untersuchungen verknüpft werden. Diese Verknüpfung basiert auf einer statistischen Analyse der Koexistenz von Termen und vergebenen Metainformationen für ein gegebenes Dokument.

Diese Entwicklung eines zusätzlichen Datenkonstrukts, des Entry Vocabulary Index, ist dabei grundsätzlich nicht zwingend erforderlich. Eine dynamische, lokale (d.h. auf eine Gruppe von potentiell relevanten Dokumenten angewendete) Lösung minimiert den Aufwand der Pflege und der stetigen Aktualisierung eines zusätzlichen Datenbestands. Da die Datenbasis des FIV stetig wächst und aktuellere Themen, mit denen sich die Beiträge befassen, ebenfalls in vielen Fällen neues, freies Vokabular mit sich bringen, wäre eine regelmäßige Neuberechnung notwendig. Schließlich ist zu erwarten, dass ein großes Interesse daran besteht, auch die neuesten Ergänzungen der Datenbank effektiv aufzufinden. Darüber hinaus wird in [Xu, Croft, 2006] beschrieben, dass zumindest für Terme aus dem Freitext eines Datenbestands ein lokaler, dem Relevance Feedback verwandter Ansatz nicht schlechter geeignet sein muss als eine globale Berechnung.

Ein weiteres Argument gegen eine globale Auswertung des paarweisen Auftretens von Termen des freien Vokabulars und Deskriptoren ist die begrenzte Anzahl von Dokumenten mit Freitexten in Form von Zusammenfassungen (rund ein Viertel aller Dokumente). Eine entsprechende Auswertung würde die Deskriptoren der mit Zusammenfassungen ausgestatteten Dokumente voraussichtlich anders in Relation zu den Termen des Freitextes stellen als es bei den Dokumenten der Fall wäre, in denen die Deskriptoren nur mit den Termen der Titel in Relation gebracht werden könnten. Würde sich eine solche globale Analyse nur auf die Zusammenfassungen beziehen, könnte nur ein Viertel der Dokumente entsprechend ausgewertet werden. Bei einer dynamischen Lösung können dagegen auch Metadaten zueinander in Relation gebracht werden, indem in einem Suchprozess anhand der bereits extrahierten Metadaten gesucht wird und weitere Deskriptoren extrahiert werden können, die besonders häufig in den gefundenen Dokumenten auftreten. Ein solcher Ansatz würde bei dem Datenbestand des FIV so gut wie alle Dokumente in einer Auswertung mit einbeziehen, da nahezu alle Dokumente mit Deskriptoren erschlossen sind.

Für das entwickelte Information Retrieval System wurden mehrere Open Source Bibliotheken verwendet. Fundament hierfür ist die Klassenbibliothek Apache Lucene in Version 1.9.1, zum Parsen und Indexieren der XML-Dateien wird Jakarta Commons Digester in Version 1.7 verwendet. Basis für die Systementwicklung bilden die an der Universität Hildesheim entwickelten Komponenten, die auf Lucene basieren und im Rahmen von CLEF erfolgreich für mehrsprachiges Retrieval eingesetzt wurden, vgl. [Hackl, Mandl, Womser-Hacker, 2005] und [Hackl, Mandl, 2006]. Zur Evaluierung des Systems kommen sowohl ein Relevanzbewertungsprogramm des Informationszentrum Sozialwissenschaften in Bonn als auch das an der Universität

[1] siehe auch: http://www.clef-campaign.org

Hildesheim nach Java portierte Programm Trec_Eval (im Original von Gerard Salton und Chris Buckley) in Version 0.7 zum Einsatz.

3 Die Indexierung

3.1 Datendateien

Die Indexierung der insgesamt 600 XML-Datendateien des Fachinformationsverbunds für Internationale Beziehungen und Länderkunde geschieht mit Hilfe des XML-Parser Jakarta Commons Digester. Es wird auf die Felder

- *file*
  (Dateiname der Datendatei)
- *id*
  (Collection/Publication/Identifier/Text)
- *title*
  (Collection/Publication/Text)
- *abstract*
  (Collection/Publication/Description/Text)
- *subject*
  (Collection/Publication/Subject/Text)
- *language*
  (Collection/Publication/Language/Text)
- *classification*
  (Collection/Publication/Classification/Text)
- *geo*
  (Collection/Publication/GeographicCoverage/Text)
- *temp*
  (Collection/Publication/TemporalCoverage/Text)

indexiert, wobei während der Indexierung der Lucene StandardAnalyzer verwendet wird. Außerdem wird eine manuell an den Datenbestand angepasste Stoppwortliste eingesetzt, die aus den besonders hochfrequent auftretenden Termen des Index mit Hilfe der Lucene Index Toolbox Luke in Version 0.6 entwickelt wurde und auch in einigen Termen an die Anfragen angepasst wurde.

Bei mehreren Einträgen in der gleichen XML-Elementebene eines Dokuments wird das entsprechende Feld des Lucene Index um alle weiteren Einträge erweitert.

3.2 Thesaurusdateien

Auch das Parsing der neun XML-Dateien des Thesaurus wird mit Hilfe von Digester realisiert, es wird auf die folgenden Felder indexiert:

- *subject*
  (Collection/Subject/Text)
- *translations*
  (Collection/Subject/Subject/Text, Typerkennung)
- *group*
  (Collection/Subject/Subject/Text, Typerkennung)
- *subGroup*
  (Collection/Subject/Subject/Text, Typerkennung)
- *connectedTerms*
  (Collection/Subject/Subject/Text, Typerkennung)

Aufgrund der im Vergleich zu den GIRT-Daten komplizierter zu parsenden XML-Architektur sowohl der Daten- als auch der Thesaurusdateien, wird beim Indexieren der Elemente unter Collection/Subject/Subject/Text im Thesaurus eine zusätzliche Methode verwendet, die auf Basis des Eintrags in dem Element entscheidet, ob es sich um eine Übersetzung, eine Gruppe, eine Untergruppe oder einen weiteren, verknüpften Begriff handelt. Diese Entscheidung basiert beispielsweise auf der Morphologie des Eintrags, der bei Gruppen z.B. einen gewissen Gruppen- oder Untergruppencode als Präfix enthält.

4 Entwicklung der einzelnen Module

Der Suchprozess verläuft im entwickelten System in einzelnen Etappen, in denen nacheinander zunächst die Suchanfrage auf sinntragende Elemente reduziert, dann die reduzierte Anfrage mit Hilfe des Thesaurus übersetzt und per Blind Relevance Feedback sowie vom Entry Vocabulary Modul erweitert wird.

Die eigentliche und abschließende Suche wird zum Ende des gesamten Prozess mit Hilfe der durch die einzelnen Schritte augmentierten Anfrage durchgeführt und das Ergebnis zu Evaluierungszwecken in ein TREC-übliches Evaluierungsdateiformat geschrieben. Der Cut-Off liegt hier bei 200 Dokumenten pro Suchanfrage.

4.1 Reduzierung der Suchanfrage auf sinntragende Elemente

Um aus der natürlichsprachlichen Anfrage die sinntragenden Elemente herauszufiltern und so eine sowohl im Umfang reduzierte als auch inhaltlich verdichtete Anfrage zu formulieren, wird im vorliegenden Retrievalsystem das Discriminator Modul eingesetzt.

Die Aufgabe des Discriminator Moduls ist es, Begriffe, die bereits beim Indexieren durch die Stoppwortliste ausgeschlossen wurden und Begriffe, die nicht während der Indexierung ausgeschlossen wurden, und besonders häufig im Index vorkommen, aus der Anfrage zu entfernen.

Hierbei ist das Ziel, dass das System hier nicht eine semantisch zentrale, freie Vokabel aus der Anfrage entfernt, nur weil sie nicht im kontrollierten Vokabular der Metadaten der einzelnen Dokumente vorkommt. Darum sucht das Discriminator Modul nach jedem einzelnen Term der Anfrage im Datenindex auf die freien Suchfelder *abstract* und *title*. Ergibt sich mindestens ein Treffer in einem der beiden Felder, verbleibt der Term in der weiter zu verwendenden Anfrage. Ergibt sich bei der Suche in den Daten bei einem Term eine Trefferliste mit mehr als 8000 Einträgen (Zahl auf Basis von Tests mit SWP-Evaluierungsanfragen festgestellt, Schwellenwert knapp über dem mit rund 7500 Nennungen am häufigsten verzeichneten Begriff „China“ von allen Begriffen der Evaluierungsanfragen), wird dieser Term im weiteren Suchprozess ignoriert. Diese Reduktion der Anfrage ist besonders für den Übersetzungsprozess im Translator Modul notwendig, da ohne eine solche Verkürzung oftmals versucht würde, sinnfreie Teile der Anfrage zu übersetzen und somit die Anfrage weiter um semantisch irrelevante Passagen zu ergänzen.

4.2 Übersetzung der Anfrage

In [Petras, 2005] wurden erfolgreiche Versuche der Übersetzung mit Hilfe eines mehrsprachigen Thesaurus, mit einem zusätzlichen Hinweis auf [Petras *et al.*, 2003], beschrieben. Entsprechend wurde versucht, auch in diesem Rahmen eine Übersetzung von Anfragetermen durch den Thesaurus zu realisieren. Da die Übersetzungsleistung in einem Retrievalsystem für eine inhaltlich spezifische Domäne zu großen Teilen von der thematischen Eignung des eingesetzten Wörterbuchs abhängt, war die Nutzung

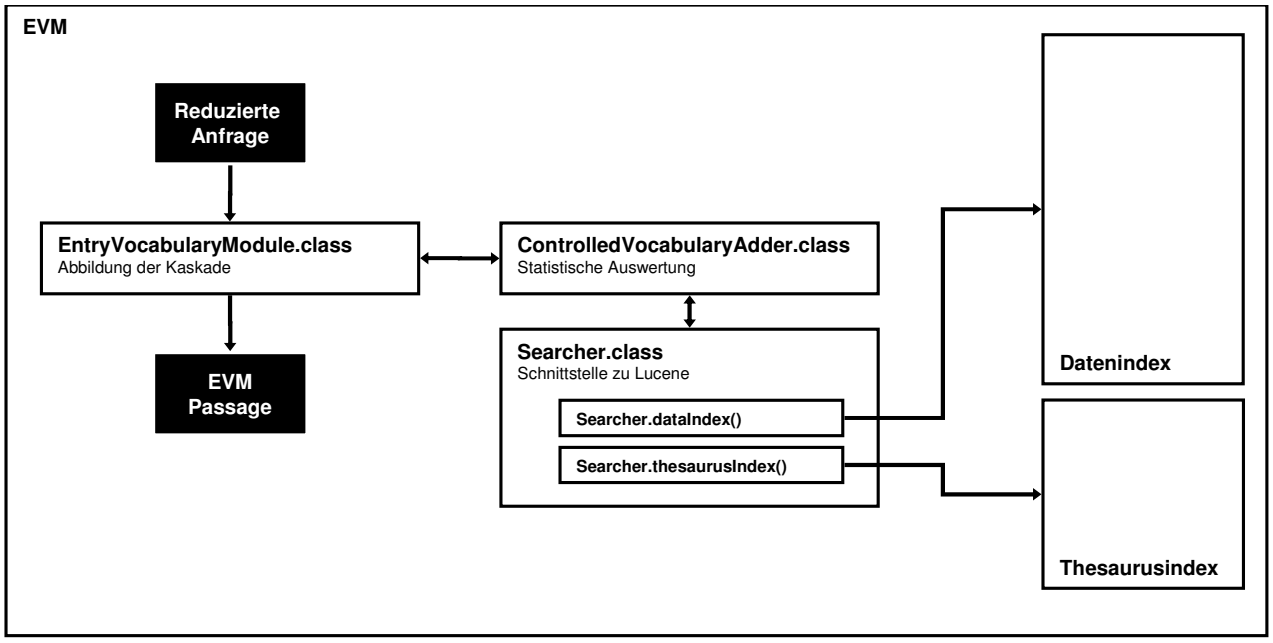

Abbildung 1: Prozessablauf des EVM

	Anfrage	Anfragefeld	Rückgabefeld	Ergebnisstring
An den Thesaurus gerichtete Anfragen:				
1.	stoppedQuery	subject	connectedterms	evmCtFromSub
2.	stoppedQuery	connectedterms	subject	evmSubFromCt
An den Datenindex gerichtete Anfragen:				
3.	stoppedQuery	title	subject	evmSubjectsFromTitleQuery
4.	stoppedQuery	abstract	subject	evmSubjectsQuery
5.	stoppedQuery	abstract	geo	evmGeoQuery
6.	evmGeoQuery	geo	subject	evmSubjectsFromGeoQuery
7.	evmSubjectsQuery	subject	classification	evmClassFromSubjectsQ.
8.	evmSubjectsQuery	subject	geo	evmGeoFromSubjectsQuery
9.	evmClassFromSubjectsQ.	classification	subject	evmSubjectsFromClassQ.

Tabelle 1: Detaillierter Verlauf der Kaskade

der Übersetzungen des vorliegenden Thesaurus naheliegend und vielversprechend.

Dem Translator Modul wird die vom Discriminator Modul reduzierte Suchanfrage übergeben. Das Übersetzungsmodul durchsuch daraufhin das Feld *subject* des Thesaurus und gibt bei einem Score von 1,0, also einer exakten Übereinstimmung, die Inhalte des Feldes *translations* zurück. Da das parallel entwickelte Modul zur Erkennung von Phrasen nicht rechtzeitig fertiggestellt werden konnte, muss sich das Translator Modul entsprechend auf termweise Übersetzungen beschränken, auch wenn bewusst ist, dass die Übersetzung von Phrasen mitunter bessere Ergebnisse erzielen kann.

Trotzdem hat eine solche Methode zur Übersetzung grundlegendes Potential: Beispielsweise lassen sich eintermige Ländernamen und Themenangaben, die zentralen Charakter für den Sinn einer Anfrage beinhalten, auf diese Art und Weise zuverlässig in Englisch, Französisch und Spanisch in der für die Datenbasis passenden Fachterminologie übersetzen.

Darüber hinaus sei darauf hingewiesen, dass der Einsatz des bereits beschriebene EVM in dem IR-System für die Datenbank des FIV einen Großteil der mehrsprachigen Retrievalleistung realisiert, da das kontrollierte Vokabular über die Dokumente aller Sprachen in einer einheitlichen Sprache verfasst ist und sich somit mit Hilfe der Deskriptoren ein grundlegendes, mehrsprachiges Retrievalverfahren realisieren lässt.

4.3 Blind Relevance Feedback

Das Blind Relevance Feedback Modul (BRF) wurde nach geringfügiger Anpassung auf den erzeugten Index aus dem System der Universität Hildesheim übernommen. Es wurde bereits in mehreren Systemen erprobt, vgl. beispielhaft [Hackl, Mandl, Womser-Hacker, 2005].

Während das im folgenden Kapitel vorgestellte EVM das lokale Feedback im Bezug auf Metadatenfelder realisieren soll, wird das BRF zusätzlich eingesetzt, um auch die Freitext-Felder *title* und *abstract* für Relevance Feedback zu nutzen. Auf diese Weise werden alle zur Verfügung stehenden Felder durch die beiden verschiedenen Varianten des Feedbacks genutzt.

Dem Blind Relevance Modul wird ebenfalls die durch den Discriminator reduzierte Anfrage übergeben. Im Rahmen der Evaluierung werden bei Einsatz des BRF pro Anfrage die 30 am besten bewerteten Dokumente untersucht und fünf Terme zur Anfragenergänzung zurückgegeben und auf die Felder *title* und *abstract* gerichtet. Die beiden Werte haben sich im Rahmen einer vorbereitenden Erprobung als vergleichsweise geeignet herausgestellt. Es

wird die Berechnungsmethode des Robertson Selection Value verwendet.

4.4 Entry Vocabulary Modul

Der Prozess dieses EVM ist in Abbildung 1 abgebildet: Zunächst wird die reduzierte Anfrage an die Klasse *EntryVocabularyModule* übergeben, dann, in den mehreren Schritten der Kaskade in der Klasse *ControlledVocabularyAdder* mit Hilfe der Klasse *Searcher* auf diverse Indexfelder angewendet und die Ergebnisse statistisch ausgewertet. Die am höchsten bewerteten Terme und Phrasen werden an die Klasse *EntryVocabularyModule* übergeben, wo alle Einzelergebnisse der Kaskadenelemente zusammengefasst und zur Ergänzung der ursprünglichen Anfrage zurückgegeben werden.

Im Falle der vorliegenden Datenbasis sind die drei für die Extraktion relevanten Felder der Datenbasis *geo*, *subject* und *classification*. Darüber hinaus werden die Felder *subject* und *connectedterms* des Thesaurus für eine Extraktion berücksichtigt. Die beiden freien Felder der Datenbasis, auf die zunächst die reduzierte Anfrage gerichtet wird sind *title* und *abstract*.

Das vorgestellte Modell wendet eine Anfragenkaskade an, die sowohl die reduzierte Anfrage, als auch aus der reduzierten Anfrage gewonnene Deskriptoren zur Gewinnung von weiteren Metadaten verwendet. Einen genauen Überblick über den Verlauf der Kaskade im evaluierten System gibt Tabelle 1. In der Tabelle werden detailliert die verwendete Anfrage, die Richtung der Anfrage auf das gegebene Feld und das ausgewertete Feld der gefundenen Dokumente sowie der zurückgegebene String aus potentiell nützlichen Deskriptoren genannt. Die Anfragenkaskade ist in der Klasse *EntryVocabularyModule* programmiert.

Im Rahmen der statistischen Auswertung wird das Suchergebnis jedes Kaskadenelements in *ControlledVocabularyAdder* untersucht. Der Umfang der Untersuchung lässt sich durch den Faktor *consideredDocs* steuern: Hier wird angegeben, wie viele der am höchsten bewerteten Dokumente in die statistische Auswertung eingehen. Über die in *consideredDocs* genannte Zahl von Dokumenten werden alle Terme und Phrasen des untersuchten Felds mit dem Score ihres Ursprungsdokuments verknüpft. Bei Mehrfachnennungen werden die Werte addiert. Über 30 Dokumente werden so z.B. im Feld *subject* bei durchschnittlich 12 Deskriptoren pro Dokument 360 Deskriptoren untersucht und ausgewertet. Abschließend werden die addierten Scores der einzelnen Deskriptoren normalisiert. Bevor die Deskriptoren zurückgegeben werden, nehmen die Parameter *cutOffScore* und *numberOfReturned* Einfluß auf die Anzahl der zurückgegebenen Dokumente. *numberOfReturned* limitiert die Anzahl der maximal zurückgegebenen Deskriptoren, falls sehr viele Deskriptoren ein höheres Ergebnis erzielt haben als in *cutOffScore* gefordert. In der Evaluierung wurden durchweg maximal sechs Terme oder Phrasen pro Kaskadenelement ergänzt.

5 Evaluierung

Im Folgenden werden die einzelnen Runs und ihre jeweiligen Unterscheidungen durch verschiedenen Parameter detailliert beschrieben. Zentrale Ansatzpunkte für die Unterscheidung der EVM-Runs sind sowohl die Wahl der Anzahl der in jedem einzelnen Kaskadenelement ausgewerteten Dokumente (30 oder 100), der auf die extrahierten Begriffe und deren Scores angewendeter Cut-Off-Score (0,3 oder 0,6) und die Anzahl der der Query hinzugefügten Terme und Phrasen (maximal sechs). Diese Parameter sind für das EVM von zentraler Bedeutung, es stellt sich die Frage, welche Einflüsse die unterschiedlichen Einstellungen auf die Retrievalleistung haben werden.

5.1 Die einzelnen Runs

Bezeichnung: SwpBase1-Nmd

- Eingesetzte Funktionen: Lucene vanilla, Stoppworte durch Indexierung
- Eingesetzte Anfragen: Eingegebene Anfrage auf die Felder *abstract* und *title*.
- Parameter des EVM: EVM nicht eingesetzt
- Globale Gewichtungen: Einfache Gewichtung der eingegebenen Anfrage

Bezeichnung: SwpBase2-Md

- Eingesetzte Funktionen: Lucene vanilla, Stoppworte durch Indexierung
- Eingesetzte Anfragen: Eingegebene Anfrage auf die Felder *abstract*, *title*, *subject* und *classification*.
- Parameter des EVM: EVM nicht eingesetzt
- Globale Gewichtungen: Einfache Gewichtung der eingegebenen Anfrage

Bezeichnung: SwpEvm1

- Eingesetzte Funktionen: Discriminator, Translator, Blind Relevance Feedback, Entry Vocabulary Modul
- Eingesetzte Anfragen: StoppedQuery (termweise angewendet auf *abstract* und *title*), RelevantTerms (termweise angewendet auf *abstract* und *title*), TranslatedQuery (termweise angewendet auf *abstract* und *title*), EvmQuery (termweise angewendet auf die Ursprungsfelder, connectedterms angewendet auf *abstract*)
- Parameter des EVM: Die maximal 30 am höchsten bewerteten Dokumente werden auf Terme und Phrasen ausgewertet. Die sechs am höchsten bewerteten Terme/Phrasen werden eingesetzt. Es gibt einen Score-Cut-Off bei 0,3 für EVM-Terme. Normalisierte Gewichtung an Termen/Phrasen entsprechend der statistischen Auswertung des EVM.
- Globale Gewichtungen: StoppedQuery: Gewichtung verzehnfacht, EvmQuery: Gewichtung einfach multipliziert mit Score-Gewicht, TranslatedQuery, RelevantTerms: Gewichtung einfach

Bezeichnung: SwpEvm2

- Eingesetzte Funktionen: Discriminator, Translator, Blind Relevance Feedback, Entry Vocabulary Modul
- Eingesetzte Anfragen: StoppedQuery (termweise angewendet auf *abstract* und *title*), RelevantTerms (termweise angewendet auf *abstract* und *title*), TranslatedQuery (termweise angewendet auf *abstract* und *title*), EvmQuery (termweise angewendet auf die Ursprungsfelder, connectedterms angewendet auf *abstract*)
- Parameter des EVM: Die maximal 100 am höchsten bewerteten Dokumente werden auf Terme und Phrasen ausgewertet. Die maximal (siehe Cut-Off)

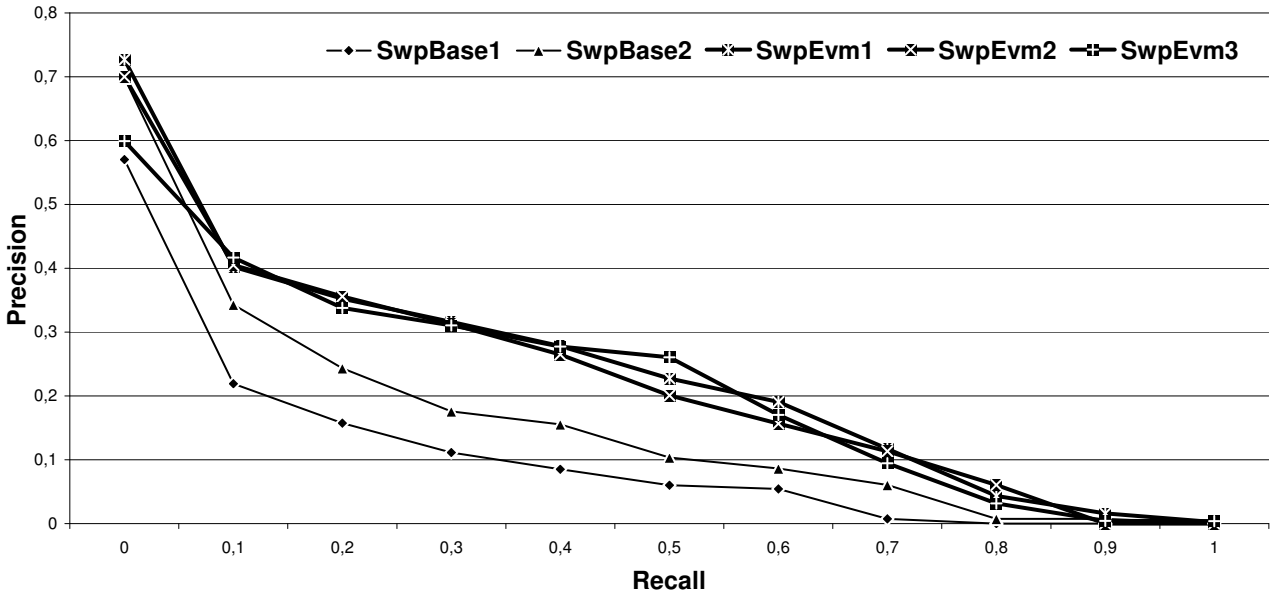

Abbildung 2: Vorläufige Ergebnisse der Evaluierung über 19 von 25 Anfragen

	BaseRun1	BaseRun2	SwpEvm1	SwpEvm2	SwpEvm3
Retrieved	3800	3800	3800	3800	3800
Relevant	1395	1395	1395	1395	1395
Rel-Ret	440	596	863	858	869
0,0	0,5705	0,6984	0,7268	0,7001	0,5992
0,1	0,2190	0,3428	0,4022	0,4049	0,4169
0,2	0,1575	0,2432	0,3525	0,3560	0,3377
0,3	0,1112	0,1759	0,3159	0,3125	0,3104
0,4	0,0851	0,1554	0,2780	0,2649	0,2775
0,5	0,0602	0,1036	0,2274	0,2008	0,2606
0,6	0,0543	0,0865	0,1907	0,1567	0,1703
0,7	0,0075	0,0604	0,1172	0,1131	0,0945
0,8	0,0000	0,0076	0,0435	0,0607	0,0315
0,9	0,0000	0,0076	0,0162	0,0000	0,0043
1,0	0,0000	0,0000	0,0022	0,0000	0,0041
Avg.Prec.	0,0865	0,1410	0,2126	0,2049	0,2070

Tabelle 2: Quantitative Ergebnisse nach Auswertung von 19 aus 25 Evaluierungsfragen

sechs am höchsten bewerteten Terme/Phrasen werden eingesetzt. Es gibt einen Score-Cut-Off bei 0,3 für EVM-Terme. Normalisierte Gewichtung an Termen/Phrasen entsprechend der statistischen Auswertung des EVM.

- Globale Gewichtungen: StoppedQuery: Gewichtung verzehnfacht, EvmQuery: Gewichtung einfach, multipliziert mit Score-Gewicht, TranslatedQuery, RelevantTerms: Gewichtung einfach

Bezeichnung: SwpEvm3

- Eingesetzte Funktionen: Discriminator, Translator, Blind Relevance Feedback, Entry Vocabulary Modul
- Eingesetzte Anfragen: StoppedQuery (termweise angewendet auf *abstract* und *title*), RelevantTerms (termweise angewendet auf *abstract* und *title*), TranslatedQuery (termweise angewendet auf *abstract* und *title*), EvmQuery (termweise angewendet auf *abstract* und *title*)
- Parameter des EVM: Die maximal 100 am höchsten bewerteten Dokumente werden auf Terme und Phrasen ausgewertet. Alle Terme/Phrasen werden eingesetzt, die über dem Score-Cut-Off von 0,6 liegen. Es wird keine spezielle Gewichtung der einzelnen Terme vorgenommen.
- Globale Gewichtungen: StoppedQuery: Gewichtung verzehnfacht, EvmQuery, TranslatedQuery, RelevantTerms: Gewichtung einfach

5.2 Der Evaluierungsprozess und erste Ergebnisse

Zur Evaluierung wurden durch die SWP 25 Evaluierungsanfragen zur Verfügung gestellt, die in ihrem Informationsgehalt, d.h. Umfang an geopolitischen und thematischen Anhaltspunkten für das IR-System, und ihrer Konkretisierung stark schwanken und somit einen hohen Anspruch an die Leistungsfähigkeit des Systems stellen. Beispielhaft können hier die folgenden zwei Fragen vorgestellt werden:

- *Frage 3: Welche Faktoren bestimmen die Beziehungen zwischen China und der EU / den einzelnen EU-Ländern?*

- *Frage 4: Welche Gefährdungen bestehen für die maritime Sicherheit in Südostasien?*

Die Evaluierung der Suchmaschine mit sämtlichen Zusatzmodulen erfolgt über einen Vergleich der Retrievalergebnisse von zwei verschiedenen BaseRuns und drei verschiedenen Evm-Runs bei denen das grundsätzliche EVM-System in diversen Parametern angepasst wurde, um den Einfluss der Parameter auf das Retrievalergebnis zu untersuchen. Zusätzlich wurde in dem Projekt das System auf den Datensatz GIRT3 evaluiert sowie auch die Leistungsunterschiede zwischen den einzelnen Modulen untersucht.

Da die Evaluierung noch andauert, kann in diesem Paper nur ein vorläufiger Überblick über das Retrievalergebnis von 19 der insgesamt 25 Evaluierungsanfragen gegen werden.

Insgesamt ist die Retrievalleistung gegenüber beiden BaseRuns deutlich angestiegen. Sowohl die Precision als auch der Recall der Evm-Runs übertrifft die Retrievalleistung, die erreicht wird, wenn die Metadaten (wie im ersten BaseRun abgebildet) nicht zum Suchvorgang hinzugezogen werden, deutlich. Im Vergleich zum zweiten BaseRun hängt die Retrievalleistung der EVM-aktivierten System allerdings deutlich von der untersuchten Query ab, trotzdem zeigt sich bei den vorliegenden Ergebnissen einen drastische Steigerung der Retrievalleistung.

6 Fazit

Da die Ergebnisse des vorgestellten IR-Systems noch nicht abschließend evaluiert und ausgewertet wurden, kann nur ein aktueller Einblick in die Entwicklungsarbeit gegeben werden.

Nach der Evaluierung von 19 aus insgesamt 25 Evaluierungsanfragen lässt sich aber vorläufig zusammenfassen, dass die EVM-basierten Runs im direkten Vergleich zu den Runs ohne entsprechende Unterstützung eine drastische Steigerung des Recalls verzeichnen können. Gleichzeitig liegt die Precision der einzelnen EVM-Runs stetig über denen der beiden BaseRuns, insgesamt lässt sich eine signifikante Steigerung der Retrievalleistung belegen.

Das dynamische EVM hat sich angesichts der ersten Evaluierungsergebnisse bewährt. In einem weiteren Schritt soll das hier vorgestellte System auch auf die GIRT Datenbasis angewandt werden, um einen besseren Vergleich mit bereits implementierten Systemen zu ermöglichen.

Literatur

[Berghaus, 2006] Benjamin Berghaus. *Mehrsprachiges Information Retrieval durch Entry Vocabulary Modul am Bespiel der Datengrundlage des Fachinformationsverbunds für Internationale Beziehungen und Länderkunde*. Magisterarbeit, Universität Hildesheim, Informationswissenschaft. 2006. erscheint.

[Buckland *et al.*, 1999] Michael Buckland, Aitao Chen, Hui-Min Chen, Youngin Kim, Byron Lam, Ray Larson, Barbara Norgard, Jacek Purat and Fredric Gey. *Mapping Entry Vocabulary to Unfamiliar Metadata*. In: Meta-Data '99 Third IEEE Meta-Data Conference, April 1999, Bethesda, USA.

[Fachinformationsverbund IBLK, 2006] World Affairs Online http://www.fiv-iblk.de *verifiziert am 30. Juli 2006*.

[Gey *et al.*, 2001] Fredric Gey, Michael Buckland, Aitao Chen and Ray Larson. *Entry vocabulary - a technology to enhance digital search*. In: Proceedings of the first international conference on Human language technology research, März 2001, San Diego, USA, S. 91-95.

[Hackl, Mandl, 2006] René Hackl, Thomas Mandl. *Bilingual Retrieval Experiments with Social Science Documents*. In: Carol Peters, Fredric Gey, Julio Gonzalo, Gareth Jones, Michael Kluck, Bernardo Magnini, Henning Müller, Maarten de Rijke. Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Berlin et al.: Springer [Lecture Notes in Computer Science 4022]

[Hackl, Mandl, Womser-Hacker, 2005] René Hackl, Thomas Mandl, Christa Womser-Hacker. *Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim*. In: Carol Peters, Paul Clough, Julio Gonzalo, Michael Kluck, Gareth Jones, Bernard Magnini: Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [Lecture Notes in Computer Science 3491] S. 165-169.

[Hellweg et al., 2001] Heiko Hellweg, Jürgen Krause, Thomas Mandl, Jutta Marx, Matthias Müller, Peter Mutschke, Robert Strötgen. *Treatment of Semantic Heterogeneity in Information Retrieval*. IZ-Arbeitsbericht Nr. 23, IZ Sozialwissenschaften, Bonn. http://www.gesis.org/Publikationen/Berichte /IZ_Arbeitsberichte/index.htm#ab23

[Kluck, 2004] Michael Kluck. *The GIRT Data in the Evaluation of CLIR Systems - from 1997 until 2003*. In: Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. S. 376-390

[Mandl, 2006] René Hackl, Thomas Mandl, Christa Womser-Hacker. *Neue Entwicklungen bei den Evaluierungsinitiativen im Information Retrieval*. Thomas Mandl, Christa Womser-Hacker (Hrsg.): Effektive Information Retrieval Verfahren in der Praxis: Proceedings Vierter Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20.7.2005. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 45] S. 117-128.

[Norgard, 1998] Barbara Norgard. *Entry Vocabulary Modules and Agents*. Technical Report

[Petras, 2005] Vivien Petras. *How One Word Can Make all the Difference - Using Subject Metadata for Automatic Query Expansion and Reformulation*. Working Notes for the CLEF 2005 Workshop, September 2005, Wien, Österreich.

[Stiftung Wissenschaft und Politik, 2006] Die Datenbasis: Inhalte. http://www.fiv-iblk.de/db/inhalte.htm *verifiziert am 30. Juli 2006*.

[Virtuelle Fachbibliothek Politikwissenschaften, 2006] Rechercheportal für die Politikwissenschaft. http://www.vifapol.de/ *verifiziert am 30. Juli 2006*.

[Xu, Croft, 2006] Query Expansion Using Local and Global Document Analysis. In: Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, S.4 - 11

[Petras, 2005] Vivien Petras. Multilingual Information Access for Text, Speech and Images. In: GIRT and the Use of Subject Metadata for Retrieval, 2005, Band 3491/2005, S.298-309

[Petras *et al.*, 2003] Vivien Petras, Natalia Perleman und Fredric Gey. Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections. In: Advances in Cross-Language Information Retrieval, 2003, S.349-362